Privacy Preservation and Inference With Minimal Mobility Information

Julián Salas[1,3][0000−0003−1787−0654] and Miguel Nunez-del-Prado[2][0000−0001−7997−1739]

[1] Internet Interdisciplinary Institute (IN3)
Universitat Oberta de Catalunya
Barcelona, Spain
jsalaspi@uoc.edu
[2] Universidad del Pacífico Av. Salaverry 2020 - Lima, Peru
m.nunezdelpradoc@up.edu.pe
[3] CYBERCAT-Center for Cybersecurity Research of Catalonia

Abstract. There is much debate about the challenge to anonymize a large amount of information obtained in big data scenarios. Besides, it is even harder considering inferences from data may be used as additional adversary knowledge. This is the case of geo-located data, where the Points of Interest (POIs) may have additional information that can be used to link them to a user's real identity. However, in most cases, when a model of the raw data is published, this processing protects up to some point the privacy of the data subjects by minimizing the published information. In this paper, we measure the privacy obtained by the minimization of the POIs published when we apply the Mobility Markov Chain (MMC) model, which extracts the most important POIs of an individual. We consider the gender inferences that an adversary may obtain from publishing the MMC model together with additional information such as the gender or age distribution of each POI, or the aggregated gender distribution of all the POIs visited by a data subject. We measure the unicity obtained after applying the MMC model, and the probability that an adversary that knows some POIs in the data before processing may be able to link them with the POIs published after the MMC model. Finally, we measure the anonymity lost when adding the gender attribute to the side knowledge of an adversary that has access to the MMC model. We test our algorithms on a real transaction database.

Keywords: Mobility Markov Chain · gender inference · geo-located data privacy · data protection regulation.

1 Introduction

Inferring demographics are useful for targeting, profiling costumers and improve services provided to them by all kind of businesses. Demographic data, together with the mobility profiles of customers, may be used for better understanding

their choices and preferences. However, revealing attributes like the home locations or the gender of users in a database may lead to re-identification, and possibly to discrimination. Hence, protecting from such inferences and finding a balance between the amount of data processed and the privacy of the individuals is an important open issue.

In this paper, we study the effects of minimizing the POIs and the data published on privacy protection. Concretely, we apply the MMC model [7], which extracts the most important POIs of an individual and guarantees data minimization, a privacy design strategy [4] that has been suggested by the General Data Protection Regulation (GDPR) from EU for improving privacy in the data processing. We study the inference of the gender attribute after applying the MMC model together with minimal additional information, then compare the anonymity (measured as the unicity) obtained by applying the MMC model with the unicity of the data before processing. We measure the anonymity lost when the gender attribute is added as side knowledge of an adversary, which uses observed POIs of an individual to try to re-identify him on the dataset.

Finally, we obtain a general formula that relates the number of raw POIs, the published POIs and the adversary's side knowledge of raw POIs, to the probability that he is able to link them to the published POIs. We use it to measure how privacy is improved by minimization of the published POIs.

The present effort is organized as follows: Section 2 describes the related works. Then, Section 3 introduces the basic concepts, while Section 4 presents the results. Finally, Section 5 concludes the paper and gives some new research avenues.

2 Related works

Different kind of datasets have been used to infer demographics. For instance, Hu *et al.* [12] use the browsing behavior to predict them. First, they propagate users age and gender to browsed pages and predict the Web pages gender and age tendency. Then, using a Bayesian framework, they predict the age and gender of users based on the tendency of the Web pages that he or she has browsed. Bi *et al.* [1] show that users traits may be predicted from search query logs by applying models developed on the Facebook Likes data.

Weinsberg *et al.* [21] show that a recommender system can infer the gender of a user with high accuracy, based uniquely on the ratings to movies provided by users, and a relatively small number of users who share their demographics. They use Movielens and Flixster dataset, in which some users have included their demographic information. Their strategy for gender obfuscation is to add fake movies to a user's record, which are strongly correlated to the opposite gender of the user. This is achieved with a random, sampled or greedy strategies. Thus, privacy protection comes from increasing the movies for all users.

Wang *et al.* [20] use Structured Neural Embedding (SNE) Model to infer age, gender and marital status. They model the possible multi labels as an eight-bit vector. Then, authors compare SNE, Pseudo Outer-Product (POP),

Singular Value Decomposition (SVD) Single, SVD-Structured, and Join Neural Embedding (JNE) algorithms for the attributes prediction. They use the BeiRen dataset, which is composed of 49 290 149 transactions over 220 828 items belonging to 1 206 379 users during 2012-2013. The obtained results in terms of weighted precision are ranging from 0.083 to 0.371.

Other works in the literature use mobility traces to infer demographic attributes. For example, Mayer *et al.* [13] studied privacy implications from metadata of mobile users. They used DBSCAN and extract the business that users have visited from telephone metadata, to re-identify the home location of individuals, assuming that these should be clustered around their home. Other algorithms for home location inference have also been studied in [3, 19] and mechanisms for protection from this inference have been provided in [18].

From the before mentioned works, we observe that home location inference is relatively straightforward when using GPS, check-ins or other real-time location data. Moreover, location data may be used for linking users among different domains [16].

Wang *et al.* [20] use non-negative matrix factorization to deduce demographic attributes based on mobile devices trajectories when connecting to access points. The authors collected data from 51 903 mobile devices for 21 days on two campuses. Authors build a three order tensor composed of mobile phones social networks (u), time (t) and location (V). Then, they associate the rank-one vector to the demographic characteristics of mobile devices users. This labelled vector is the input of an SVM and Logistic Regression classification algorithm. They used 90% as training and the rest as test obtaining a precision between 69.9% and 72.2%.

Zang *et al.* [23] propose a framework for demographic inference. In detail, they use the Mobile Data Challenge dataset to infer demographics of people. Authors reconstruct a graph representing social ties among individuals of the dataset using a multilabel regression technique. Besides, authors label individuals with their respective demographic attributes. Therefore, to infer socio-demographic attributes from the labelled dataset, they use the average of the prediction made by ten different classification algorithms tree-based, linear and kernel-based to infer individuals gender. Authors obtained an accuracy between 69.23% and 92.30%.

Zhong *et al.* [24] propose a location to profile (L2P) framework for inferring demographic attributes of online users. This inference includes both gender and age from location check-ins considering the temporality and spatiality. The former estimates at what times the individuals develop their activities such as transportation or shopping on their POIs. The later guesses where are those activities carried out (location of POIs). Also, they try to know the semantics of the POIs (*i.e.,* category, price, and range of a restaurant), which can be enriched with customer reviews or comments.

Regarding privacy protection, several methods have been developed for trajectory data, cf. [5]. While, others have measured the unicity of mobility and transaction data (even when it is coarsened) for privacy protection [14],[15]. How-

ever, none of them has assessed the effect of minimizing the published records on the privacy of individuals. Thus, we follow an approach that goes in a different direction from previous works, instead of generating an increasing amount of attributes and collecting more information, we minimize the amount of information collected, even at the expense of precision, considering privacy to be a priority. We infer demographics and test the unicity of the data, to measure the privacy preserved by applying the MMC model from [7]. We also measure the loss in anonymity when considering the additional knowledge of gender or age attributes.

Finally, we give a general formula for measuring the privacy obtained by the minimization of the POIs published. Data minimization is one of the diverse challenges for big data privacy protection [17], and is consistent with recent policy recommendations (e.g., GDPR law from EU) for guaranteeing privacy by design in data collection and publishing [4].

3 Background

In the present section, we introduce the mobility model and the adversary side knowledge used for the inference attacks. It is worth noting that we select the MMC mobility model [7] due it has been use to perform different inference attacks, such as POI extraction [9], next whereabouts prediction [8], and de-anonimization [10].

3.1 Mobility Markov Chain

A *Mobility Markov Chain* (MMC) [7] models the mobility behavior of an individual as a discrete stochastic process in which the probability of moving to a state (*i.e.*, a point of interest) depends only on the previously visited state and the probability distribution on the transitions between states. More precisely, a MMC is composed of:

- A set of states $P = \{p_1, \dots p_n\}$, in which each state is a frequent POI (ranked by decreasing order of importance), with the exception of the last state p_n that corresponds to the set made of all infrequented POIs. POIs are usually learned by running a clustering algorithm on the mobility traces of an individual. These states generally have an intrinsic semantic meaning and therefore semantic labels such as "home" or "work" can often be inferred and attached to them.
- Transitions, such as $t_{i,j}$, represent the probability of moving from state p_i to state p_j. A transition from one state to itself is possible if the individual has a non-null probability from moving from one state to an occasional location before coming back to this state. For instance, an individual can leave home to go to the pharmacy and then come back to his home. In this example, it is likely that the pharmacy will not be extracted as a POI by the clustering algorithm, unless the individual visits this place on a regular basis.

Note that many mobility models relying on a Markov chains have been proposed in the past [7], including the use of hidden Markov models for performing inference attacks [22]. In a nutshell, building a MMC is a two step process. During the first phase, a clustering algorithm is run to extract the POIs from the mobility traces. We use the clustering algorithm called Density Joinable cluster (DJ-Cluster) that was used in the study of Gambs *et al.* [7], however, other clustering algorithms are possible. In the second phase, the transitions between those POIs are computed.

DJ-Cluster takes as input a trail of mobility traces and 3 parameters: the minimal number of points $MinPts$ needed to create a cluster, the maximum radius r of the circle within which the points of a cluster should be contained and a distance d at which neighboring clusters are merged into a single one. DJ Cluster works in three phases. During the first phase, which corresponds to a pre-processing step, all the mobility traces in which the individual is in movement (*i.e.*, whose speed is above some small predefined value), as well as subsequent static redundant traces, are removed. As a result, only static traces are kept. The second step consists in the clustering itself: all remaining traces are processed in order to extract clusters that have at least $MinPts$ points within a radius r of the center of the cluster. Finally, the last phase merges all clusters that have a trace in common or whose centroids are within d distance of each other.

Once the POIs (*i.e.*, the states of the Markov chain) are identified, the probabilities of the transitions between states can be computed. To realize this, the trail of mobility traces is examined by chronological order and each mobility trace is tagged with a label that is either the number identifying a particular state of the MMC or the value "unknown". Finally, when all the mobility traces have been labeled, the transitions between states are counted and normalized by the total number of transitions in order to obtain the probabilities of each transition. The MMC is represented as a transition matrix of size $n \times n$, the rows and columns correspond to states of the MMC while the value of each cell is the probability of the associated transition between the corresponding states.

The *predictability* [8] is a theoretical measure quantifying how predictable is the mobility of an individual based on his MMC model. The predictability $Pred$ of a particular user u corresponds to the sum of the product between each element $\pi_{i,u}$ of the stationary vector computed from the MMC of user u:

$$Pred(u) = \sum_{i=1}^{n_u} (\pi_{i,u} \times p_{max\_out}(i,u)), \tag{1}$$

in which $\pi_{i,u}$ is the probability of being in a particular state (for n_u, the total number of states of the MMC of user u) and $p_{max\_out}(i,u)$ represents the maximum outgoing probability leaving from the i^{th} state.

In general, the *(Shannon) entropy* is a measure of uncertainty regarding the output of a random variable. In the context of mobility, the entropy of a user quantifies the spatial uncertainty about the exact location of a user. Considering a particular user u, we can compute his entropy by applying the

following formula:

$$H(u) = -\sum_{i=1}^{n_u} p_{i,u} \log_2 p_{i,u}, \tag{2}$$

in which p_i, u represents the probability of user u to be located in his i^{th} POI while n_u corresponds to the total number of POIs characterizing his mobility.

In Table 1 we show the information generated using the MMC model from a transactions dataset. As we can see each row shows a Point of Interest (in this case represented by the Business ID), a user identifier (Name), his/her age, gender, home district, user's entropy, user's predictability, and a stationary probability corresponding to that POI.

User ID	Age	Gender	District	Entropy	Predictability	Business ID	Stat probability
Bob	49	M	SAN ISIDRO	1.53	0.89	10907215	0.33
Bob	49	M	SAN ISIDRO	1.53	0.89	10010695	0.45
Bob	49	M	SAN ISIDRO	1.53	0.89	10012805	0.22

Table 1. MMC properties example.

3.2 Adversary Model

For defining the adversary knowledge, we introduce the four pieces of information an adversary may have for performing the gender inference. First, the *MMC properties* from Table 1.

The second piece of information is the *aggregated probability*, which considers the POIs visited by each user and the gender distribution by POI. Then, we aggregate the probabilities of all the POIs that a user has visited to obtain the gender probability for each user. This information is computed based on the aforementioned *MMC properties*. Table 2a shows an example of this information.

Finally, the *male/female count* is the gender proportion by POI as shown in Table 2b and the *male/female count per age* is the gender proportion by POI and age range as depicted in Table 2c.

Based on the before mentioned pieces of information, we have built five different scenarios with distinct cases. Each scenario and case denote the use of a combination of one or two pieces of information to represent the scenario attack by the adversary as introduced in Table 3. For instance, in Scenario three the adversary uses the *MMC properties* as well as the *aggregated probability* as prior information to perform the attack.

In the following section, we detail specific cases of these different scenarios to perform the gender inference attack.

a) Aggregated probability

User ID	Female probability
Bob	0.35
Alice	0.8
Eve	0.82

b) Male/female count

Business ID	Male	Female
00128826	523	497
00155626	179	234
00112326	139	156

c) Male/female count per age example

Business ID	18-25		25-30		30-40		40-50		50-60		60	
	M	F	M	F	M	F	M	F	M	F	M	F
00136626	35	30	3	7	24	20	14	17	32	37	6	6
00136624	135	230	13	17	32	25	17	19	37	39	16	26
00136628	53	34	31	74	21	20	37	57	23	77	11	34

Table 2. Adversary knowledge examples.

Scenario	MMC properties	Aggregated prob	M/F count	Age
1	✓	-	-	-
2	✓	-	✓	-
3	✓	✓	-	-
4	✓	-	-	✓
5	-	✓	-	-

Table 3. Adversary *a priori* knowledge. Where *Age* is the male and female count by age.

4 Results

In the present section, we perform some experiments for inferring the gender of bank users based on their transactions and also carry out unicity tests concerning their POIs.

4.1 Experiments

In the present effort, we use 65 millions of banking transactions dataset, which is composed of 1) pseudonym of the user as ID; 2) age; 3) gender; 4) Merchant Category Code (MCC)[4]; 5) the timestamp of the transaction; 6) the number of monetary units spent; 7) quantity of transactions; and, 8) the district ID of the transaction. The dataset was gathered from June 2016 to May 2017 in Peru.

First of all, we need to find a suitable classification algorithm for implementing the inference attack. In the present work, we use four different classification algorithms, namely Gradient Boosted Regression Trees [6], Extremely Randomized Trees [2], Discrete and Real AdaBoost [11]. Figure 1 shows the precision when inferring the gender using the *MMC properties* as the input of the different classification algorithms per number of estimators, and we observe that Gradient Boosted Regression Trees performs the best with 199 estimators. Therefore, we use this algorithm for further experiments.

[4] VISA Merchant Category Classification (MCC) codes directory.

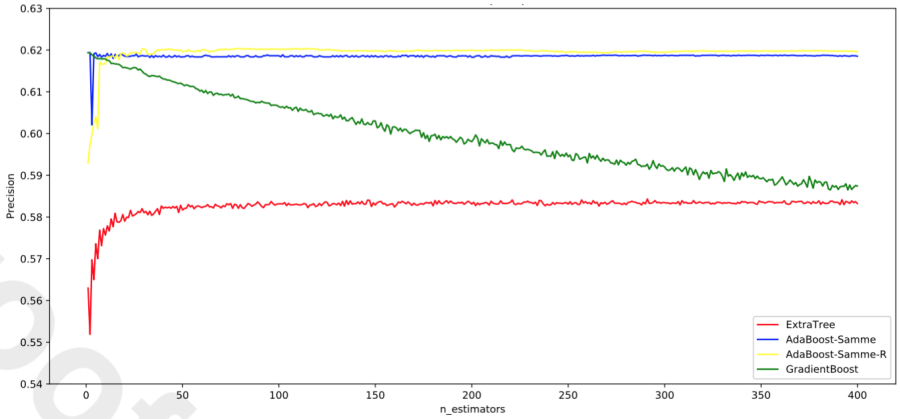

Fig. 1. Evaluation of the Gradient Boosted Regression Trees (*GradientBoost*), Extremely Randomized Trees (*ExtraTree*), Discrete AdaBoost (*AdaBoost Samme*), and Real AdaBoost (*AdaBoost Samme R*) for inferring users' gender.

Since we have settled the classification algorithm, the adversary performs attacks in different scenarios and cases, as shown in Table 4. We observed seven different scenarios for gender inference. The adversary uses either 10-fold cross-validation or Grid search for the rest. For instance, in Scenario one Case one *S1C1*, the adversary uses Gradient Boosted Regression Trees (*GB*) algorithm over the *MMC properties* to infer gender. We observe a precision of 0.56, which is poor. In *S2C1*, the adversary combines two pieces of information, the *MMC properties* and the *male/female count per business*. Therefore, *GB* takes as input the *MMC properties* for training. Then, when the classifier estimates whether a subject is female if the probability of the prediction is under the threshold (60%), the adversary looks at the supplementary knowledge (*i.e.*, the male/female count) to infer the gender. Thus, the adversary uses a voting system where each POI votes for male or female based on the number of male or female visitors. In the same spirit, scenarios *S2C2* and *S3C1* utilize *GB* taking as input the *MMC properties* for training but they vary for the decision making. In the former, the adversary chooses whether a person is male if the sum of all males visitors in the POIs of his MMC model is bigger than the number of the females' counterparts. The latter follows the same logic, but instead of the *male/female count*, it employs the *aggregated probability* to perform the inference. Therefore, for scenarios *S2C1*, *S2C2*, and *S3C1*, we have obtained 57.65%, 57.8%, and 59.07% as best precisions, respectively. Finally, scenario *S4C1* adopts the same process as scenario *S2C2*, but it uses the *male/female count* by age to decide whether a given user is male if the classification probability is under a threshold of 60% obtaining 59.01%. It is worth noting that thresholds have been chosen empirically.

Concerning the scenarios *S3C2* and *S5C1*, they use Logistic Regression (*LR*) algorithm with Grid Search to fine tune the parameters of the regressor. In scenario *S5C1*, the adversary applies the LR directly over the *aggregated prob-*

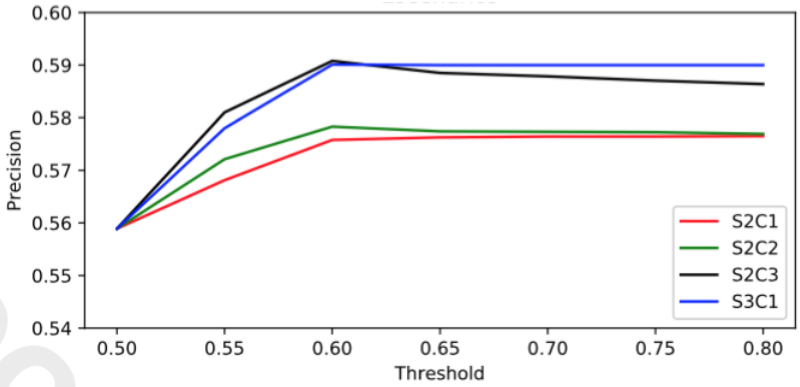

Fig. 2. Precision obtained depending on the threshold.

code	Scenario	Case	Algorithm	Training	Precision	Threshold
S1C1	1	1	GB	10-fold cross validation	0.56	-
S2C1	2	1	GB	10-fold cross validation	0.5765	0.6
S2C2	2	2	GB	10-fold cross validation	0.578	0.6
S2C3	3	1	GB	10-fold cross validation	0.5907	0.6
S5C1	3	2	LR	Grid search	0.63	-
S3C1	4	1	GB	10-fold cross validation	0.5901	0.6
S4C1	5	1	LR	Grid search	0.62	-

Table 4. Attack scenarios. Where GB = Gradient Boosted Regression Trees and LR = Logistic Regression.

abilities information reaching a precision of 62%. While, in scenario *S3C2* the adversary utilizes also LR algorithm over the merge of two pieces of information such as *MMC properties* and *aggregated probabilities* getting the best precision of all (*i.e.,* 63%).

4.2 Unicity tests

In this section, we study the unicity of the transaction records after we apply the MMC model. This information is used in [14] to show that the mobility traces are highly unique even when their time and location are coarsened, and thus, they are susceptible to re-identification.

We will prove that publishing a smaller number of POIs (as in the MMC) even when they are more relevant, will decrease the possibility of finding unique records in our published data considerably. This implies that linking unique records with additional outside information is harder when using the MMC.

First, we test the unicity of the MMC, considering two, three, four, and five random POIs as the adversary knowledge, for each of the records in the MMC. Observe that most of the records have only 2 POIs in the MMC, while in the original database they may have many more, as we can see in Table 5.

	Original	MMC
Min.	2	2
1st Qu.	28	2
Median	51	2
Mean	67.1	2.4
3st Qu.	89	3
Max.	674	13

Table 5. Statistics of number of POIs per user after MMC.

We depict the unicity in Figure 3 by sampling $p \in \{2, 3, 4, 5\}$ random POIs for each user and counting the number of records that contain the same p points. If there is only one user with such p points, it can be uniquely identified by this adversary's knowledge. We compare how the unicity increases if the adversary knows also the gender attribute. For example, the unicity increases from 23.8% to 35.7% for an adversary that knows 2 POIs and learns the gender attribute.

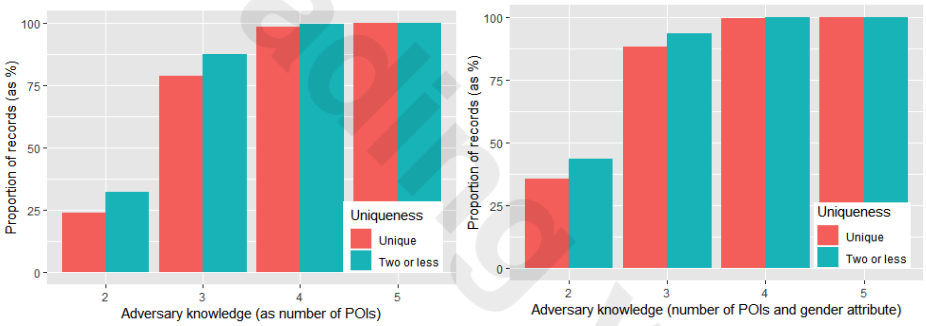

Fig. 3. unicity of records after obtaining MMC, depending on the number of POIs that an adversary knows and gender attribute.

Privacy Obtained by Minimization.- As we have previously explained, the MMC model does not publish all the POIs of an individual but only a few of the most relevant. So, we would like to measure how much privacy is gained by publishing less POIs than the original POIs for each individual.

For this, we consider an adversary that knows x random POIs of the original data of a user and tries to link them to his published data on the MMC.

First, we calculate the probability that x random POIs of adversary knowledge in the original data contain two specific POIs of the MMC of a given user.

Then, we consider how many POIs an adversary must know to have the certainty (with probabilities 0.25, 0.5, 0.75, 0.95 and 1) that such two POIs are published in the MMC. This is depicted in Figure 4.

We can see that in average an adversary has to know 34.38 POIs to have the certainty up to 25% that such side knowledge of a user will contain the two POIs of the user published in the MMC. If the adversary wants to have a certainty of 50% he must have 48.2 POIs of side knowledge. For 75% he needs 58.81 POIs, for 95% needs 66.06 POIs, and for 100% certainty needs to know all the POIs that in average are 67.27, see Figure 4.

Another way of interpreting these results is that if an adversary knows 34 POIs of a user which has 2 POIs in the MMC model, he may be able to uniquely identify such user in the data with a probability 0.057 = 0.25*0.23, which is the probability that the 34 POIs contain that user's 2 POIs published in the MMC and that such 2 POIs are unique. Thus, we can see that the minimization of the published POIs in the MMC is effective for privacy protection.

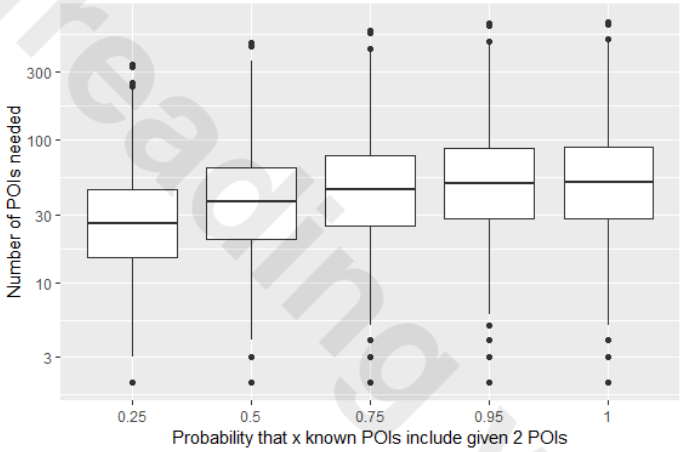

Fig. 4. Probability that given user with x POIs in the original data, contain 2 specific POIs of the MMC of a given user.

Generalization.- We consider that an individual has n POIs, there have only been published $m \geq 2$ and the adversary knows r of the original n POIs. We give a closed formula for the probability (denoted by $Pr(Y = p)$)that an adversary who knows r of the original POIs has exactly p of them among the published POIs.

Therefore, we consider that $m < n$ data points have been published and the adversary knows $r \leq n$ of the original POIs.

We know that all the possible subsets with r points are $\binom{n}{r}$. If we assume that there are exactly p of the r POIs of the adversary that are part of the published POIs, then there are $r - p$ adversary POIs in the remaining $n - m$ not published POIs. Hence, there are $\binom{m}{p}$ ways in which their p adversary POIs

belong to the published POIs and $\binom{n-m}{r-p}$ possible ways in which the remaining adversary POIs are distributed among the non-published POIs.

Therefore, in general, the probability that an adversary who knows r real POIs, gets to know precisely p of the m published POIs is:

$$Pr(Y = p) = \frac{\binom{m}{p} * \binom{n-m}{r-p}}{\binom{n}{r}} \tag{3}$$

We make an example, considering a user from our dataset with 67 POIs (average number of POIS in our dataset), and assume that we have published only seven of them. In Figure 5, we use equation (3) to calculate the different probabilities that an adversary knows at least two, three, four, five, six or all the published POIs, depending on the number of POIs that he might have as auxiliary (previous) knowledge.

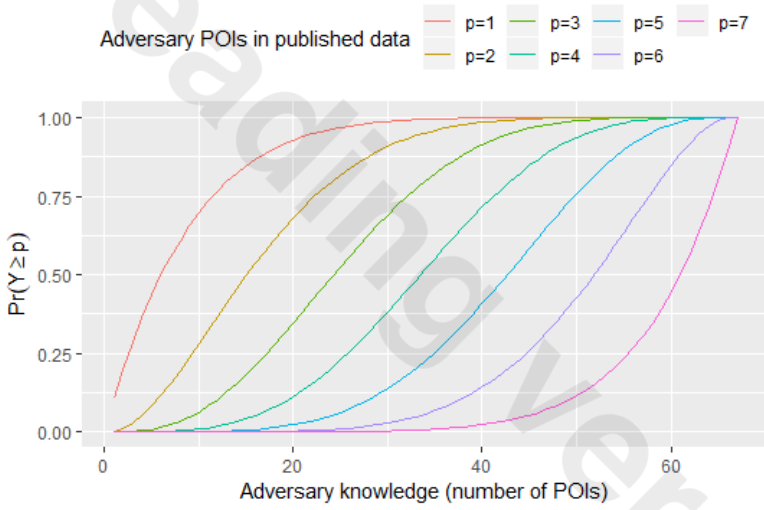

Fig. 5. Example of user with 7 published POIs and 67 POIs before publishing.

It is worth noting that an adversary needs more data points r to know at least p POIs.

5 Conclusions

In this paper, we have studied the effect that minimization has on the privacy of the data published after applying the MMC model. We considered the gender inference that can be carried out using the MMC model, gender and age distributions by POI. We have used a banking transactions dataset processed

with the MMC model, together with the Male/Female and age distributions by POI to make inferences with Gradient Boosted Regression Trees, Extremely Randomized Trees and Logistic Regression.

We have investigated which are the best algorithms for inference and made comparisons of the different combinations of side knowledge to infer the gender attribute. We obtained that the Logistic Regression with the aggregated knowledge of the gender distributions of all the POIs visited by an individual gives the best precision (63%).

We have calculated the unicity of records published with the MMC and measured how it decreases if an adversary also infers the gender attribute. In the worst case, knowing the gender attribute besides 2 POIs would allow him to identify 11.9% more records uniquely. Finally, we have measured the capability of an adversary to link several POIs of side knowledge from raw data to a record after applying the MMC. We have shown that an adversary must know several points, to have a reasonable probability of linking his knowledge to a unique record, for example, an adversary must know in average half of the raw POIs of an individual to link this knowledge to the two corresponding POIs published after the MMC model with 0.25 certainty. Hence, in that case, the probability of finding a unique record in the MMC model data is 0.05. This shows the effectivity of data minimization for protecting privacy obtained from the MMC model.

Possible future work is to repeat our study with Call Detail Record (CDR) data, to obtain a general formula for the probability of linking the raw data points to any given amount of POIs published and to test the effects of minimization of data publishing on the inference of other attributes.

Acknowledgement

This work was supported by the Spanish Government, in part under Grant RTI2018-095094-B-C22 "CONSENT", and in part under Grant TIN2014-57364-C2-2-R "SMARTGLACIS."

References

1. Bi, B., Shokouhi, M., Kosinski, M., Graepel, T.: Inferring the demographics of search users: Social data meets search queries. In: Proceedings of the 22Nd International Conference on World Wide Web. pp. 131–140. WWW '13, New York, NY, USA (2013)
2. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
3. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: User movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1082–1090. KDD '11, New York, NY, USA (2011)
4. Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J.H., Métayer, D.L., Tirtea, R., Schiffner, S.: Privacy and data protection by design - from policy to engineering. Tech. rep., ENISA (2015)

5. Fiore, M., Katsikouli, P., Zavou, E., Cunche, M., Fessant, F., Hello, D.L., Aivodji, U.M., Olivier, B., Quertier, T., Stanica, R.: Privacy of trajectory micro-data : a survey. CoRR **abs/1903.12211** (2019)
6. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
7. Gambs, S., Killijian, M.O., Núñez del Prado Cortez, M.: Show me how you move and i will tell you who you are. Trans. Data Privacy **4**(2), 103–126 (Aug 2011)
8. Gambs, S., Killijian, M.O., Núñez del Prado Cortez, M.: Next place prediction using mobility Markov chains. In: Proceedings of the First Workshop on Measurement, Privacy, and Mobility. vol. 3, pp. 1–6. Bern, Switzerland (April 2012)
9. Gambs, S., Killijian, M.O., del Prado Cortez, M.N.: Gepeto: a geoprivacy-enhancing toolkit. In: 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops. pp. 1071–1076. IEEE (2010)
10. Gambs, S., Killijian, M.O., del Prado Cortez, M.N.: De-anonymization attack on geolocated data. Journal of Computer and System Sciences **80**(8), 1597–1614 (2014)
11. Hastie, T., Rosset, S., Zhu, J., Zou, H.: Multi-class adaboost. Statistics and its Interface **2**(3), 349–360 (2009)
12. Hu, J., Zeng, H.J., Li, H., Niu, C., Chen, Z.: Demographic prediction based on user's browsing behavior. In: Proceedings of the 16th International Conference on World Wide Web. pp. 151–160. WWW '07, ACM, New York, NY, USA (2007)
13. Mayer, J., Mutchler, P., Mitchell, J.C.: Evaluating the privacy properties of telephone metadata. Proceedings of the National Academy of Sciences **113**(20), 5536–5541 (2016)
14. de Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: The privacy bounds of human mobility. Scientific Reports **3** (2013)
15. de Montjoye, Y.A., Radaelli, L., Singh, V.K., Pentland, A.S.: Unique in the shopping mall: On the reidentifiability of credit card metadata. Science **347**(6221), 536–539 (2015). https://doi.org/10.1126/science.1256297
16. Riederer, C., Kim, Y., Chaintreau, A., Korula, N., Lattanzi, S.: Linking users across domains with location data: Theory and validation. In: Proceedings of the 25th International Conference on World Wide Web. pp. 707–719. WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2016)
17. Salas, J., Domingo-Ferrer, J.: Some basics on privacy techniques, anonymization and their big data challenges. Mathematics in Computer Science **12**(3), 263–274 (Sep 2018)
18. Salas, J., Megías, D., Torra, V.: Swapmob: Swapping trajectories for mobility anonymization. In: Domingo-Ferrer, J., Montes, F. (eds.) Privacy in Statistical Databases. pp. 331–346. Springer International Publishing, Cham (2018)
19. Scellato, S., Noulas, A., Lambiotte, R., Mascolo, C.: Socio-spatial properties of online location-based social networks. In: Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011 (2011)
20. Wang, P., Guo, J., Lan, Y., Xu, J., Cheng, X.: Your cart tells you: Inferring demographic attributes from purchase data. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 173–182. ACM (2016)
21. Weinsberg, U., Bhagat, S., Ioannidis, S., Taft, N.: Blurme: Inferring and obfuscating user gender based on ratings. In: Proceedings of the Sixth ACM Conference on Recommender Systems. pp. 195–202. RecSys '12, New York, NY, USA (2012)

22. Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., Aberer, K.: Semitri: A framework for semantic annotation of heterogeneous trajectories. In: Proceedings of the 14th International Conference on Extending Database Technology. pp. 259–270. EDBT/ICDT '11, ACM, New York, NY, USA (2011)
23. Zhong, E., Tan, B., Mo, K., Yang, Q.: User demographics prediction based on mobile data. Pervasive Mob. Comput. **9**(6), 823–837 (Dec 2013)
24. Zhong, Y., Yuan, N.J., Zhong, W., Zhang, F., Xie, X.: You are where you go: Inferring demographic attributes from location check-ins. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 295–304. WSDM '15, ACM, New York, NY, USA (2015)